# Measuring the incremental information value of documents

Th.P. van der Weide and P. van Bommel

Radboud University Nijmegen, The Netherlands

**Abstract**

The *incremental searcher satisfaction model* for Information Retrieval has been introduced to capture the incremental information value of documents. In this paper, from various cognitive perspectives, searcher requirements are derived in terms of the increment function. Different approaches for the construction of increment functions are identified, such as the *individual* and the *collective* approach. Translating the requirements to similarity functions leads to the so-called base similarity features and the monotonicity similarity features. We show that most concrete similarity functions in IR, such as Inclusion, Jaccard's, Dice's, and Cosine coefficient, and some other approaches to similarity functions, possess the base similarity features. The Inclusion coefficient also satisfies the monotonicity features.

## 1 Introduction

Finding relevant documents no longer seems to be the major challenge of state-of-the-art search engines. Were recall and precision major concerns in the early days of their existence, trying to convey information rather than just data seems to be a major concern nowadays. Offering a long list of documents in order of their relevancy score is known to be a too simple interface. Several approaches have been attempted to improve on this. A central place is the construction of an overview which is understandable and may be used as a base for further searching. Another key issue is a presentation metaphor.

From research as reported in [3] the most important reasons for searching information are:

1. looking for new developments
2. having a concrete information need
3. exploring a new field of interest

The 1st and 3rd activity have some analogy and are different from the 2nd activity. Addressing the 1st and 3rd activity, we assume two typical searcher moods. A searcher may be in an explorative mood, open for new information. This searcher will benefit from variety rather than extensiveness. Having explored an area of interest, the searcher may want to exploit it by selecting a coverage of this area of interest. In this exploitative mood, the searcher will benefit from a concise overview. During the search session a searcher may remain in one of these moods or alternate between exploration and exploitation successively (see figure 1). We will recognize several types of searchers, each having their own balance between exploration and exploitation.

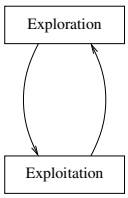


Figure 1. Balancing between variety and conciseness

In our approach, the underlying metaphor is a retrieval session, which starts upon entering a query. We will assume an anonymous searcher. As a consequence, when starting the retrieval session, the profile of that searcher is empty. In this paper we restrict ourselves to recording documents that actually have been retrieved by the searcher, or otherwise may be assumed to be familiar to the searcher. As a consequence, the user profile will consist of a sequence of (presented) documents. The order in which documents are retrieved is important when this order may be interpreted as a manifestation of a drifting information need.

In this paper we study retrieval sessions from the standpoint of retrieval models. Traditional retrieval models restrict themselves to estimating relevancy of documents in the context of a single user request. The *incremental searcher satisfaction model* is a conditional approach to relevance estimation. Document relevancy is considered in the light of a document profile ([22]). The conditional relevance function is referred to as the *increment function*. The intention of this approach is maximizing search support (cf. [4], [7]).

The embedding of incremental information content within the framework of economics has been addressed in [21] where it is argued from an economical point of view that the standard paradigm of information retrieval to present documents in order of relevance is not sufficient as it does not take into account the incremental value of the documents already viewed. As in [22], they argue that there is no point in offering a document twice, and that offering a document similar to those earlier in the list, adds little value to those already examined. Another conclusion of [21] is that clustering is good when it maximizes the difference (the incremental information content) across clusters. An overview of related work on recommender systems is found in [1].

Different approaches for the construction of increment functions are identified (see figure 1). The idea is to define several ways to compare a document with the document profile. Two approaches are studied in-depth: the *individual* and the *collective* approach. The requirements posed by these approaches are defined within an axiomatic framework. We show that collective increment functions have a strict nature, posing more requirements than individual increment functions. The principles underlying the incremental model are further examined by confronting above-mentioned approaches with existing similarity measures.

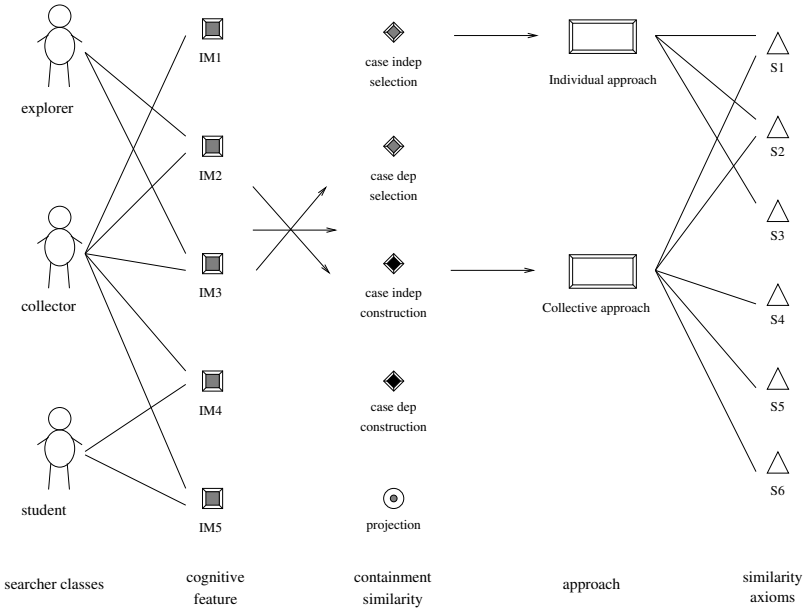


Figure 2: Levels of refinement

The incremental model can be used in combination with other techniques in this area, such as document ranking techniques (e.g. [19]) and techniques for visualizing relevancy (see e.g. [11]). Although we focus on the kernel of IR relevancy treatment and pay little attention to the user interface, we propose the incremental model to be embedded within systems having interaction features especially suited for IR applications (see e.g. [6], [5]). Furthermore, incremental relevancy can be applied in the area of document summarization. For details about incremental summarization see [9], where a linear combination of two similarity functions is used, one for quickly selecting a set of documents, which is more closely investigated by a second, more accurate similarity function.

In [9] the maximal marginal relevance function is introduced as a mechanism to estimate differential relevance of documents. The motivation behind the maximal marginal relevance function has some resemblance to the increment func-

| **searcher type** | repetition | growth | effectiveness | independence | exclusion |
|---|---|---|---|---|---|
| globe trotter | - | × | × | - | - |
| student | - | - | - | × | × |
| collector | × | × | × | × | × |

Table 1: Cognitive identities characterized

tion, but they are rather different in their properties. See [12] for more background on this topic.

The organization of the paper is as follows. In section 2, we distinguish some searcher classes, and model them in terms of cognitive identities. These identities are formally described in terms of the incremental model. In section 3 focus is on realizations of the incremental model, and make an in-depth study of similarity between a collection of documents and a single document. Collection-object similarity is related to object-object similarity functions. In section 4 we evaluate the familiar object-object similarity functions as a base for collection-object similarity. Two approaches to collection-object similarity, referred to as the individual and the collective approach, will be elaborated to illustrate the complete coverage approach. Finally, in section 5 we present conclusions and further research.

# 2 Searcher model

The Information Retrieval paradigm is about a person (physical or not) having a need for information, and a document collection from which this need is to be satisfied. In this paper we focus on isolated retrieval sessions, where knowledge transfer between retrieval sessions and collaboration between different searchers is not considered. Typically no single document (also referred to as information object) from the collection can completely satisfy (cover) the need of the searcher. A search engine will shift this problem to the searcher and simply offers the documents in order of estimated relevance. Our approach goes a step further focussing on a complete coverage of the information need.

## 2.1 General setup of increment functions

In the incremental searcher satisfaction model ([22]), or *incremental model* for short, it is assumed that the need for documents is influenced by what the searcher already has retrieved from the archive. This can be modelled as a function

$$I : \wp(\mathcal{O}) \times \mathcal{O} \mapsto [0, 1]$$

$I(S, x)$ is interpreted as the residual need for document $x$ after the set $S$ has been presented to the searcher. The function $I$ is also referred to as the *increment function*.

A special case occurs when a document is presented without any previously presented documents. This is the case at the start of a retrieval session. The initial increment value $I(\varnothing, x)$ is also referred to as the (initial) document need (denoted as $N(x)$).

The set $S$ can also be interpreted as the personal knowledge of the searcher (sometimes also called a user profile) during a retrieval session. The set $S$ of already presented documents then acts as a *mini-profile* of the searcher.

The incremental model is especially useful for (very) dynamic and distributed archives, such as the World Wide Web. Firstly, as the increment function allows for real-time calculation. This is in contrast with approaches that try to cluster the retrieval result before presenting the clusters to the searcher. Secondly, for distributed archives recall is not useful as a measure for retrieval quality. We rather use a quality measure based on total searcher satisfaction (the coverage problem), bypassing the need to have global knowledge of the collections involved (see also [22]).

## 2.2 Cognitive Settings

As described in the introduction, different searchers may have a different context of the information need, for example different tasks and motivations. We will be interested in a cognitive characterization of the person having the need for information. We introduce a framework in which the cognitive settings of various searcher classes can be formalized in terms of search behavior. We discuss some examples.

**The globe trotter** The first searcher class we consider is the *globe trotter*, examining a particular field of interest in order to find out sufficient details. In terms of the search process, an globe trotter is seen as a searcher trying to cover some topic of interest, without really being interested in completeness. Experiencing new sensations is the incentive of this searcher.

**The student** Next we consider the cognitive setting of the *student*. A student is a searcher who is trying to get acquainted with some topic. The topic is not stable, reading a document might draw the student's attention to a new area of interest. Reading an information object a second time may be profitable, especially when documents read in between have contributed knowledge that enables the student to learn more in a second reading pass.

**The collector** A rather different searcher class is the *collector*. A collector is a searcher wishing to collect information objects with respect to some topic. It is not profitable to have an object more than once. The collector tries to make the collection complete.

In this paper, we introduce a number of properties to characterize peculiarities of searcher classes:

- The property of *repetition* describes the effect of repetition.
- The property of *growth* is about the effect of growing knowledge.
- *Effectiveness* focuses on informational dependencies between documents.
- The property of *independence* describes the relation between independent documents.
- The *exclusion* property relates informational dependence and independence.

These properties will be introduced in terms of the framework from next section. In table 1 we characterize the cognitive searcher identities in terms of these properties.

#### 2.2.1 Repetition

The first cognitive feature we consider deals with *repetition*. The effect of repetition may be rather diverse. For example, a globetrotter (a special kind of globe trotter) will not appreciate visiting a region twice. After visiting the fjords from Norway, this globetrotter will not be interested in a second trip to this location. In terms of the incremental model, this is expressed as:

$$\text{IM1} \quad \textit{Repetition:} \qquad x \in S \Rightarrow I(S, x) = 0$$

For this same reason it does not make sense when a search engine offers a document twice in the result list of a query. Note however that there are also many situation that handle repetition differently. For example, after drinking a glass of beer, some people might feel a desire for another glass.

#### 2.2.2 Growing knowledge

The second cognitive feature deals with the effect of growing knowledge. In many cases the following holds: the more you experience, the more you know. In the context of information retrieval, the consequence of this rule is that providing a document leads to (partial) satisfaction of the information need of that searcher. So, this feature expresses that the more you know, the less you need. This is formulated as follows:

$$\text{IM2} \quad \textit{Growth:} \qquad S \subseteq T \Rightarrow I(S, x) \geq I(T, x)$$

Note that the feature of growth does not necessarily always hold. For example, visiting Italy may give the globetrotter a stronger wish to visit Greece.

The cognitive features IM1 and IM2 are tailored to a classical information retrieval environment, in which no distinction is made between alternative searcher types. The motivation for these features is the assumption that presenting documents has a satisfying (non-increasing) effect on the document need. In table 1 we see that the globe trotter as well as the collector have this property.

Some immediate consequences of the two basic cognitive features are:

1. IM1 $\vdash\ I(\{x\}, x) = 0$
2. IM2 $\vdash\ I(S, x) \leq N(x)$

The first consequence immediately follows from cognitive feature IM1. On the other hand, IM1 can be derived from $I(\{x\}, x) = 0$ combined with IM2. The second is an immediate consequence of cognitive feature IM2.

### 2.2.3 Effective knowledge

In this section we introduce a third cognitive feature based on effective knowledge. This feature is expressed in terms of information containment for documents. Information containment is used as a basis for aboutness in the context of matching information objects with queries ([8]). In terms of the incremental model, the information containment relation is defined as:

$$x \preceq_I y \ \equiv\ I(\{y\}, x) = 0$$

where $x \preceq_I y$ is verbalized as: *the information in $x$ is contained within $y$*, in the context of the information need represented by $I$. In the sequel, we will omit the index $I$, and denote information containment as $\preceq$. We will also use the generalized notation $x \preceq S$ to denote $I(S, x) = 0$.

The effect on $x$ of presenting $y$ carries over to more complex situations:

**Lemma 2.1** IM2 $\vdash\ x \preceq y \iff \forall_S\,[x \preceq S \cup \{y\}]$

Next we isolate the effect of presenting a single document.

**Lemma 2.2**
IM2 $\vdash\ I(S \cup \{y\}, x) = I(S, y) + I(\{y\}, x) \Rightarrow y \preceq S$

If the information in document $x$ is contained within $y$, then presenting document $y$ eliminates the need for document $x$:

**Lemma 2.3** IM2 $\vdash\ x \preceq y \wedge y \in S \Rightarrow x \preceq S$

**Lemma 2.4** $\forall_S\,[I(S, x) \leq I(S, y)] \Rightarrow x \preceq y$

Irrelevant documents (i.e. $N(x) = 0$) do not contain any information that is relevant for the information need of the searcher. Such documents thus can be seen as empty-information objects. Irrelevant documents have special properties:

**Lemma 2.5** IM1 $\vdash N(x) = 0 \Rightarrow x \preceq y$

In cases where cognitive feature IM1 holds, it directly follows that the relation $\preceq$ is reflexive.

**Lemma 2.6** IM1 $\vdash x \preceq x$

We now consider the possibility that the containment relation can be transitive, as this makes the containment relation a partial order on documents. This partial order plays a vital role in the reasoning process within logical models of Information Retrieval (see [14] or [10]).

Transitivity is enforced by the next cognitive feature, dealing with *effectiveness*:

$$\text{IM3} \quad \textit{Effectiveness:} \quad x \preceq y \wedge y \preceq z \Rightarrow x \preceq z$$

We also consider the following stronger form, which guarantees feature IM3, but not vice versa:

$$\text{IM3a} \quad \textit{Effective Growth:} \quad x \preceq y \Rightarrow I(S, x) \leq I(S, y)$$

**Lemma 2.7** IM3a $\Rightarrow$ IM3

So, if the information from document $x$ is contained within $y$, then document $x$ can not be more informative than document $y$. In order to explain IM3a, consider the following question: Suppose all attractions a visitor may experience in Finland seem also to be available in Argentina. So a globe trotter, looking for new places, will find Finland less attractive than Argentina, independent of the visiting history of this trotter.

An immediate corollary of IM3a is that subdocuments can not be more relevant than superdocuments:

**Lemma 2.8** $x \preceq y \Rightarrow N(x) \leq N(y)$.

Information containment is, generally, not a symmetric relation between documents: different documents may mutually contain each others information. As a consequence, documents $x$ and $y$ are considered *equally informative*, denoted as $x \approx y$, if:

$$x \approx y \;\equiv\; x \preceq y \wedge y \preceq x$$

If two documents are equally informative, then under no circumstance, one of those documents can add something new to the other.

**Lemma 2.9** $x \approx y \Rightarrow I(S \cup \{y\}, x) = 0$

From IM3 it follows that the relation $\approx$ is an equivalence relation for documents. If one document of an equivalence class is found to be relevant for some query, then the other documents from that class are equally relevant. The notion of information preclusion can be used to further distinguish between the documents within an equivalence class.

### 2.2.4 Independent knowledge

In this section we introduce a cognitive feature related to independent knowledge. This feature is expressed in terms of the *not-about* relation. Besides similarity which basically aims at aboutness, the not-about relation is essential as well when reasoning about information retrieval (see e.g. [23]).

For a given retrieval situation, modelled by increment function $I$, a document $y$ can be considered to be not about document $x$, denoted as $x \rfloor_I y$, if:

$$x \rfloor_I y \;\equiv\; I(\{y\}, x) = I(\varnothing, x)$$

So, the relation $x \rfloor_I y$ expresses that presenting document $y$ does not influence the need for document $x$. Although important in a general context, the index $I$ will be omitted in the rest of this paper.

We start by noting that irrelevant documents have a special place. In a specific retrieval situation, they do not contain any relevant information. In that sense, irrelevant documents do not handle about anything. As a consequence, presenting such a document can not have any effect on the need for any other document:

**Lemma 2.10** $N(x) = 0 \Rightarrow x \rfloor y$

The nature of the not-about relation is laid down in the following cognitive feature. This feature deals with *independence*:

*Suppose you have visited Argentina and we offer you a trip to Finland. Now if for your decision about Finland, your earlier trip to Argentina is entirely irrelevant, what would that mean? Would this mean that going to Finland is not affected by Argentina in all future situations, independent of the countries you will visit?*

If this is the case, we say that your way of travelling conforms to the law of independence:

IM4 *Independence:* $x \rfloor y \Rightarrow I(S \cup \{y\}, x) = I(S, x)$

This cognitive feature expresses that the not-about relation is not affected by presenting more documents. If presenting a set $S$ of documents does not have any effect on the need for a document $x$, then all documents $y$ from $S$ are not about $x$:

**Lemma 2.11** $I(S, x) = N(x) \wedge y \in S \Rightarrow x \rfloor y$

For relevant documents $x$, the relations $x \preceq y$ and $x \rfloor y$ exclude each other. In other words, if $x$ is not about $y$, then the information of $x$ cannot be contained within $y$:

**Lemma 2.12** If $N(x) > 0$, then $x \rfloor y \Rightarrow x \not\preceq y$.

### 2.2.5 Exclusive knowledge

Next we consider a final cognitive feature in which the not-about relation is combined with the containment relation. This feature deals with exclusion as follows:

*Suppose you have visited Mexico and you decide that a trip to Russia would not be sufficiently interesting to you. Rather, you are considering a trip to Turkey and that for your decision in this matter, your earlier trip to Mexico is entirely irrelevant. Now what more can we say about your possible trip to Turkey?*

Is it valid to claim that for your decision about Turkey, your earlier trip to Russia is irrelevant as well? We then would say that your way of travelling conforms to the law of exclusion:

$$\text{IM5} \quad \textit{Exclusion:} \quad x \rfloor y \wedge z \preceq y \Rightarrow x \rfloor z$$

So, if document $x$ is not about $y$, and the information of document $z$ is contained within $y$, then obviously $x$ is also not about $z$. After having introduced the requirements for increment functions, we will present concrete functions in the next section.

# 3 Fundamentals of increment functions

In this section we present some concrete definitions for increment functions. For this purpose, we also consider similarity functions. We show how an increment function may be easily added to an existing IR situation in which some measure *Sim* for similarity is available. The relevance score of documents may be obtained as $Rel_q(x) = Sim(\chi(q), \chi(x))$, where $\chi(x)$ is a representation of the contents of document $x$ (for example as a set of keywords or as a document vector). We

assume $\chi(q)$ is a similar representation of the query $q$. We will overload the function $Sim$, and use $Sim(x, y) = Sim(\chi(x), \chi(y))$.

The similarity function is assumed to return a value from $[0, 1]$, where 1 is interpreted as most similar, while 0 is the lowest level for similarity. A similarity function has to satisfy the following condition (see [13]):

**S1.** *maximal similarity:* $Sim(A, A) = 1$

which states that equivalent objects are most similar. Each similarity function in fact may be seen as the extension of some equivalence relation ([18]).

The similarity function is not assumed to be a symmetric function (see [20]), and is also not required to satisfy a transitivity condition or an equivalent of the triangular inequality as is used in e.g. distance functions (see figure 3).

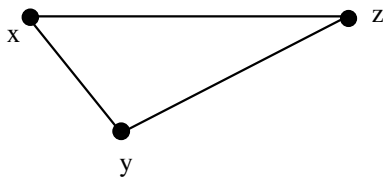


Figure 3. $Sim(x, z) \geq Sim(x, y) + Sim(y, z)$?

When $Sim(A, B) = 1$ then $A$ and $B$ are most similar, they are considered to be indiscernible from each other. This is denoted as $\mathsf{Identical}(A, B)$. Being indiscernible usually is not a transitive relation (illustrated via the famous coffee example in [16], considering a series of cups of coffee with slightly increasing amount of sugar). Note that S1 may be rewritten as:

**S1.** *maximal similarity:* $\mathsf{Identical}(A, A)$

The other extreme situation is when $A$ and $B$ are least similar: $Sim(A, B) = 0$. This is denoted as $\mathsf{Orthogonal}(A, B)$.

This section is organized as follows. In section 3.1, we continue our framework by extending the notion of similarity between documents, to similarity between a document and a collection. We describe basic techniques for exploration and exploitation here. The aim of this paper is reflected in the first possibility, which we have termed *containment similarity*. This similarity will be treated in the context of reductional approaches to containment similarity in section 3.2, whereas an approach based on projection is discussed in section 3.3. A further confrontation with elementary similarity axioms is centered around an *individual approach* in section 3.4 and a *collective approach* in section 3.5. Related work is found in [12].

## 3.1 Object-set similarity

### 3.1.1 Containment similarity

The function $SetSim(S, x)$ evaluates the similarity between a set $S$ of documents and a single document $x$ as a value from $[0, 1]$. We will use the term *containment similarity* for such a similarity function. There will be no requirements in advance for containment similarity functions. We will not even require the function $SetSim(\{y\}, x)$ to be a similarity measure for documents.

In literature, comparing a group of items with an individual has been studied in the context of fuzzy set membership. Using the relevancy score of a document as its degree of membership. This way, the retrieval result can be seen as a fuzzy set of documents.

If the function $SetSim(S, x)$ measures the containment similarity between set $S$ and document $x$, then $1 - SetSim(S, x)$ is a measure for the dissimilarity between $S$ and $x$.

To be able, during a retrieval sessions, to estimate what is new and what not, we use the searcher profile $S$ as a basis. Then the increment function should indicate, whether a document $x$ yields sufficient new information compared to this profile. As we consider profile $S$ as a set of documents assumed to be known to the searcher, we may interpret $1 - SetSim(S, x)$ as the degree in which document $x$ is unknown to the searcher. This outcome is scaled into the interval $[0, Rel_q(x)]$ (see section 2.2.1). This is typically expressed by degrading the a-priori relevance score $Rel_q(x)$ with the dissimilarity of $x$ from $S$. This leads to the *explorative increment function*, defined as:

$$I(S, x) = Rel_q(x)\,(1 - SetSim(S, x))$$

For explorative increment functions we recognize the following basic properties for information containment:

**Lemma 3.1** Let $x$ be a relevant document (i.e., $N(x) > 0$), then

- $x \subseteq y \iff SetSim(\{y\}, x) = 1$
- $x \rfloor y \iff SetSim(\{y\}, x) = 0$

### 3.1.2 Incremental coverage

Besides finding new information, search may be directed towards building an *overview* of a topic for which $S$ is an exemplary description. The searcher may wish to find a new document $x$ that contains most of the relevant information from $S$. In this case, the containment similarity function may be used to calculate the coverage coefficient of documents, also referred to as the *incremental coverage function*:

$$C(S, x) = N(x)\,SetSim(S, x)$$

The coverage coefficient may be interpreted as the conceptual distance between an individu and a group.

During a retrieval session, a single user may switch between these two types of information quest (see figure 1). For example, in a first explorative phase the aim could be to find a sufficient number of documents with new information. In this phase, broadness is a guideline, recall is not really important. During the exploitative phase the aim is switched to find documents with overview information. During this phase, other factors (such as document cost) may play a role to get a concise coverage. Note that after exploitation, it is possible to restart exploration, and so forth.

## 3.2 Containment similarity by reduction

Comparing an individu with a group is not a trivial task. A main reason is that the yardstick for comparison may be very diverse. For example, for a stamp collector the baseline is similarity with individual stamps in the collection ($S$). For a knowledge collector, however, the base for comparison is the knowledge level obtained (from $S$).

Rather than computing $SetSim(S, x)$ directly, we consider an *indirect* computation based on the document similarity function $Sim(y, x)$, where $y$ and $x$ are both documents (document characterization), that we assume to be available.

In order to apply document similarity, the collection $S$ has to be *reduced* to a single representative $y = Reduce_q(S, x)$, usually referred to as a *centroid* of $S$. So this leads us to the following approach:

$$SetSim(S, x) = Sim(x, Reduce_q(S, x))$$

Several approaches can be taken to find a representative for a set of documents. These can be categorized according to several criteria. We will restrict ourselves to the following criteria:

1. Is a centroid a *primus inter pares*, or: $Reduce_q(S, x) \in S$? The positive case, reduction by selection, is also referred to as the *individual* approach. Typically, the selected document is considered to be the main result of the search process so far. In the other case, reduction has the nature of *construction*, composing a (virtual) document with some special representation. This is also referred to as the *collective* approach.

2. Is the centroid case-dependent? The reduction of set $S$ may be guided by $x$, the document to be compared. This will be useful in applications where the nature of similarity is finding a look-a-like.

These two criteria lead to four different situations, which we will discuss below.

### 3.2.1 Case independent selection

The goal of case independent selection is to find a best general-purpose representative of a group. In the context of retrieval, the relevance score can be used as general-purpose comparison measure. Let $y = Reduce_q(S, x)$, then the reduction function should satisfy

$$y \in S \wedge \forall_{z \in S} \left[ Rel_q(z) \leq Rel_q(y) \right]$$

Informational performance is another option. In that case the reduction should satisfy:

$$y \in S \wedge \forall_{z \in S} \left[ \frac{Rel_q(z)}{Cost(z)} \leq \frac{Rel_q(y)}{Cost(y)} \right]$$

where $Cost(x)$ is the cost associated with a document, for example, the length of that document.

### 3.2.2 Case dependent selection

In this case, the reduction should select from profile $S$ the best look-a-like of $x$. An example of this reduction by selection would be to define $Reduce_q(S, x)$ as a document in $S$ with maximum similarity with $x$:

$$Reduce_q(S, x) \in \left\{ y \in S \;\middle|\; \forall_{z \in S} \left[ Sim(z, x) \leq Sim(y, x) \right] \right\}$$

There may be several possibilities for choosing the reduction $y$ in the above definition. Since the actual choice is irrelevant for the resulting *SetSim* score, we may use the following equivalent definition:

$$SetSim(S, x) = \max \left\{ Sim(x, y) \;\middle|\; y \in S \right\}$$

In section 3.4 this approach is elaborated in more detail. Instead of taking maximal similarity, it is also possible to use average or minimal similarity. Clearly, the appropriate choice here depends on the aim of the retrieval function.

### 3.2.3 Case independent construction

Quite a different situation arises, if the searcher wants to consider a general property of the documents in $S$ as a base for comparison with the new document $x$. Then the searcher does not choose a specific document, but considers properties of all documents together. Now $y \notin S$ and if reduction does not depend on $x$, this is reduction by construction using union (or e.g. average) of documents in $S$:

$$Reduce_q(S, x) = \cup S$$

In section 3.5 this approach is elaborated in more detail.

### 3.2.4 Case dependent construction

Comparing the new document $x$ with an overview property $y \notin S$ of the documents in $S$ may in some cases depend on $x$. A searcher may request this when (a) reduction by selection is too much focussed on a single known document, and (b) independent reduction by construction is not sufficiently focussed on specific known documents. Then, dependent reduction by construction provides a balance. An example of this is based on a bandwidth $B(x, S, \delta) \subseteq S$ where $\delta$ is a measure for the width. This bandwidth can be defined as follows:

$$B(x, S, \delta) = \{ z \in S \mid | Rel_q(z) - Rel_q(x)| \leq \delta \}$$

Now $Reduce_q(S, x)$ may be union (or e.g. average or even intersection) as follows:

$$Reduce_q(S, x) = \cup B(x, S, \delta)$$

It is evident that intersection makes only sense for sufficiently small $\delta$.

## 3.3 Containment similarity by projection

Computing similarity $SetSim(S, x)$ by reduction is based on the reduction of the set $S$ to a single element, followed by the application of the regular $Sim$ function. In this section we consider an alternative approach based on *projection* rather than reduction. This approach is aiming at noise reduction.

The idea behind projection is as follows. On the one hand, we have a document characterization $\chi(x)$ in terms of a set $\mathcal{I}$ of descriptors. On the other hand, we have a set of document characterizations $S$. Now in $x$ we focus on some descriptor $i \in \mathcal{I}$. In order to present the same focus in $S$, this set has to be focussed on a local property in which $i$ is reflected. This local property is obtained by projecting $S$ onto $i$:

$$\pi_i(S) = \{ y \in S \mid i \in y \}$$

It is evident that as a result of projection, we ignore information form the characterization of $x$. This naturally leads to some form of *partial similarity between $S$ and $x$ via $i \in \mathcal{I}$*. A basic partial similarity is counting the hits in the projection of $S$:

$$PartSim(S, i) = |\pi_i(S)|$$

In analogy with reduction as presented in the previous section, *SetSim* can be based on average, maximum, and other properties emerging in partial similarity. Collection/document similarity based on average projection is defined as follows:

$$SetSim(S, x) = \frac{1}{|\chi(x)|} \Sigma_{i \in \chi(x)}\, PartSim(S, i)$$

Another typical application is to restrict to the query under consideration $q$:

$$SetSim(S, x) = PartSim(S, q)$$

## 3.4 The individual approach

In the individual approach, the similarity between a document and a (non-empty) set of documents is measured as the maximal similarity between the document and any instance of this set. In this case, the function $Reduce_q$ selects from $S$ the document most similar to $x$.

$$Ind(S, x) = \max\left\{ Sim(\chi(x), \chi(y)) \;\middle|\; y \in S \right\}$$

A focussed version of this function is:

$$Ind(S, x) = \max\left\{ Sim(\chi(x) \cap \chi(q), \chi(y) \cap \chi(q)) \;\middle|\; y \in S \right\}$$

Furthermore, $Ind(\varnothing, x) = 0$. The expression $Ind(S, x)$ provides the maximal similarity between document $x$ and any of the elements from $S$ of previously presented documents. This results in the following definition for the function *SetSim*:

**Lemma 3.2** $Ind(\{y\}, x) = Sim(\chi(x), \chi(y))$

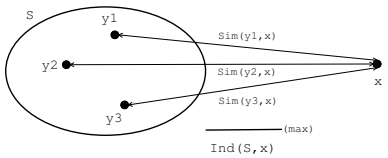


Figure 4. The individual approach

The resulting increment function is denoted as $I_i$ (see figure 3.4). Thus $I_i(S, x)$ gives the fraction of the need $N(x)$ for document $x$ not yet being covered by any previously presented document from $S$. Consequently, for two documents bringing an equal quantity of new information, the more relevant one is displayed before the less relevant one, as one would expect. Otherwise, the most exotic (and therefore probably highly surprising) documents would be presented before relevant ones.

The following conditions for similarity functions, are sufficient express the cognitive features IM1,.., IM5:

**S2.** $\mathsf{Identical}(A, B) \wedge \mathsf{Identical}(B, C) \Rightarrow \mathsf{Identical}(A, C)$

**S3.** $\mathsf{Orthogonal}(A, B) \wedge \mathsf{Identical}(C, B) \Rightarrow \mathsf{Orthogonal}(A, C)$

Note that the transitivity requirement S2 is a rather strong requirement for a similarity function.

**Lemma 3.3** $I_i(\varnothing, x) = N(x)$

**Lemma 3.4** The cognitive features IM1,.., IM5 correspond in the individual approach, both in the unfocussed as the focussed case, as follows to requirements on the base similarity function:

1. IM1 is a consequence of S1
2. IM2 is a direct consequence of the definition of *Ind*
3. IM3a is a consequence of S2
4. IM4 is a direct consequence of the definition of *Ind*
5. IM5 is a consequence of S3

## 3.5 The collective approach

In the collective approach, a new document $x$ is compared to a set $S$ of previously presented documents by comparing the characterization of $x$ with a summary of all presented material from $S$. This summary is constructed by accumulating the individual document characterizations using some operator $\cup$. The summary $\sigma(S)$ of the set $S$ is defined as follows:

$$\sigma(S) = \cup_{y \in S} \chi(y)$$

As a consequence, empty summary is $\sigma(\varnothing) = \varnothing$ and extension of summary is given by $\sigma(S \cup \{x\}) = \sigma(S) \cup \chi(x)$. The similarity between a document $x$ and a set $S$ of documents then is defined as

$$Col(S, x) = Sim(\chi(x), \sigma(S))$$

The expression $Col(S, x)$ provides the degree document $x$ is covered by the total of information provided by the elements from $S$ of previously presented documents. The collective increment function is denoted as $I_c$.

**Lemma 3.5** $Col(\{y\}, x) = Sim(\chi(x), \chi(y))$

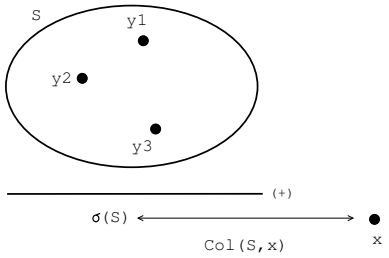


Figure 5. The collective approach

We need $I_c$ to have the basic property of increment functions $I_c(\varnothing, x) = N(x)$ mentioned in section 2.1. In the collective approach this property holds if similarity with the empty set is impossible:

**S4.** $A \neq \varnothing \Rightarrow \mathsf{Orthogonal}(A, \varnothing)$

**S5.** $Sim(X, A) \leq Sim(X, A \cup B)$

**S6.** $\mathsf{Orthogonal}(A, B) \Rightarrow Sim(A, S \cup B) = Sim(A, S)$

**Lemma 3.6** $S4 \Rightarrow I_i(\varnothing, x) = N(x)$

Next we consider the question under what conditions the cognitive features IM1 to IM5 are satisfied in the collective approach.

**Lemma 3.7** The cognitive features IM1,.., IM5 correspond in the collective approach as follows with requirements on the base similarity function:

1. IM1 is a consequence of S1 and S5
2. IM2 is a consequence of S5
3. IM3a is a consequence of S2
4. IM4 is a consequence of S6
5. IM5 is a consequence of S3

# 4 Similarity functions

In this section several instances of increment functions are considered. This is done by choosing specific similarity functions as an instantiation of the generic function *Sim* used in section 3. Each similarity function is evaluated with respect to the similarity features, making a distinction between the discrete case (where the similarity measure compares different sets) and the weighted case. In the weighted case, the intersection and union operator for characterizations can be defined as via one of the following strategies:

1. straightforward:

$$\begin{aligned} (A \cap B)_i &= A_i \cdot B_i \\ (A \cup B)_i &= A_i + B_i - A_i B_i \end{aligned}$$

2. fuzzy:

$$\begin{aligned} (A \cap B)_i &= \min(A_i, B_i) \\ (A \cup B)_i &= \max(A_i, B_i) \end{aligned}$$

Note that if a similarity feature holds for the weighted case then it also holds for the discrete case, as the discrete case results from the weighted case by restricting weights to 0 and 1. Furthermore, a counter example in the discrete case is also a counter example for the weighted case.

In the light of similarity features as introduced in the precious section, we will consider some well-known similarity functions such as Inclusion coefficient, Overlap coefficient, Jaccard's coefficient, Dice's coefficient, and Cosine coefficient, as they are found in the literature (see e.g. [17]). All these coefficients are more or less similar in their way to measure the commonality between two objects, but have different strategies to normalize the amount of commonality. We will see that all similarity functions agree in their handling of orthogonality.

Similarity functions are applied in many areas. They are used to express the degree in which two objects are found to be similar, usually on a $[0, 1]$ scale. Formally, a similarity function is introduced as follows (see [13]):

**Definition 4.1**

*A similarity function Sim on a class $\mathcal{D}$ in a weight class $\mathcal{W}$ is a mapping $Sim : \mathcal{D} \times \mathcal{D} \to \mathcal{W}$ such that:*

$$Sim(x, x) = 1_{\mathcal{W}}$$

*where $\mathcal{W}$ is a totally ordered set with having $1_{\mathcal{W}}$ as its maximal weight.*

We will restrict ourselves to weight class $\mathcal{W} = [0, 1]$. The similarity function is called *reflexive* when this function is symmetric, otherwise the similarity function is called *directed*. Note that similarity will be the result of both positive and negative contributions.

Before discussing similarity functions, we note that the similarity features can be grouped as follows. The first group, the base similarity features, covers identicality and orthogonality:

**S1.** $\mathsf{Identical}(A, A)$

**S2.** $\mathsf{Identical}(A, B) \wedge \mathsf{Identical}(B, C) \Rightarrow \mathsf{Identical}(A, C)$

**S3.** $\mathsf{Orthogonal}(A, B) \wedge \mathsf{Identical}(C, B) \Rightarrow \mathsf{Orthogonal}(A, C)$

**S4.** $A \neq \varnothing \Rightarrow \mathsf{Orthogonal}(A, \varnothing)$

The second group covers the monotonicity of similarity functions:

**S5.** $Sim(X, A) \leq Sim(X, A \cup B)$

**S6.** $\mathsf{Orthogonal}(X, B) \Rightarrow Sim(X, A) = Sim(X, A \cup B)$

Note that we have rewritten S6 slightly to emphasize its relation with S5.

## 4.1 Inclusion coefficient

We first consider the Inclusion coefficient for similarity. This coefficient normalizes the amount of overlap $A \cap B$ with the size of $A$. It is given by:

$$\mathsf{Incl}(A,B) = \frac{|A \cap B|}{|A|} = \frac{\sum_i \min(a_i, b_i)}{\sum_i a_i}$$

(denoting component $i$ of $A$ resp. $B$ as $a_i$ resp. $b_i$) in case $A$ nonempty. Furthermore $\mathsf{Incl}(\varnothing, \varnothing) = 1$, and $\mathsf{Incl}(\varnothing, B) = 0$ for non-empty $B$. As a consequence:

1. $\mathsf{Identical}(A,B) \iff \sum_i (a_i - \min(a_i, b_i)) = 0 \iff \forall_i [a_i \leq b_i]$. This latter condition is also denoted as $A \leq B$.
2. $\mathsf{Orthogonal}(A,B) \iff \sum_i (A \cap B)_i = 0 \iff \forall_i [(A \cap B)_i = 0] \iff \forall_i [\min(a_i, b_i) = 0]$. This latter condition is also denoted as $A \perp B$.

We will prove the base similarity features for the weighted case. By restricting weights to $\{0, 1\}$, we find their validity for the set-oriented case.

1. Reflexivity and transitivity of the relation $\mathsf{Identical}$ are direct consequences of the corresponding properties for the relation $\leq$ on objects.
2. Assume $\mathsf{Orthogonal}(A,B)$ and $\mathsf{Identical}(C,B)$. Then for all $i$ we have $a_i b_i = 0$ and $c_i \leq b_i$, from which $a_i c_i = 0$ can be concluded, and thus $\mathsf{Orthogonal}(A,C)$.
3. $A \neq \varnothing \Rightarrow \mathsf{Orthogonal}(A, \varnothing)$ is obvious.

The monotonicity similarity features are also valid:

1. The validity of feature S5 is a consequence of the property $\min(x, a) \leq \min(x, \max(a, b))$, and thus $\mathsf{Incl}(X, A) \leq \mathsf{Incl}(X, A \cup B)$.
2. Similarity feature S6 is a direct consequence of the property $\min(x, b) = 0 \Rightarrow \min(x, a) = \min(x, \max(a, b))$, and thus $\mathsf{Orthogonal}(X, B) \Rightarrow \mathsf{Incl}(A, S \cup B) = \mathsf{Incl}(A, S)$.

## 4.2 Jaccard's coefficient

Next, we consider Jaccard's similarity coefficient. This coefficient normalizes intersection $A \cap B$ with the corresponding union. The straightforward generalization from the set-oriented to a weighted version does not work:

$$\mathsf{Jacc}(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{\sum_i a_i b_i}{\sum_i a_i + \sum_i b_i - \sum_i a_i b_i}$$

In this case, two object are indiscernible when:

$\mathsf{Identical}(A,B) \iff \sum_i a_i b_i = \sum_i a_i + \sum_i b_i - \sum_i a_i b_i \iff \sum_i a_i(1-b_i) + \sum_i (1-a_i)b_i = 0 \iff \forall_i [a_i \neq 0 \Rightarrow b_i = 1] \wedge \forall_i [b_i \neq 0 \Rightarrow a_i = 1] \iff \forall_i [a_i = b_i \wedge a_i, b_i \in \{0,1\}]$

which means that indiscernibility is restricted to proper sets only. For the fuzzy logic approach we get:

$$\mathsf{Jacc}(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{\sum_i \min(a_i, b_i)}{\sum_i \max(a_i, b_i)}$$

in case both $A$ and $B$ are nonempty. If either $A$ or $B$ is empty, we have $\mathsf{Jacc}(A,B) = 0$. Finally, $\mathsf{Jacc}(\varnothing, \varnothing) = 1$. This leads to:

1. $\mathsf{Identical}(A,B) \iff \sum_i \min(a_i, b_i) = \sum_i \max(a_i, b_i) \iff \forall_i [a_i = b_i] \iff A = B$.
2. $\mathsf{Orthogonal}(A,B) \iff A \perp B$

The validity of the base similarity features is by similar arguments as for the Inclusion coefficient. The monotonicity similarity features do not hold for Jaccard's coefficient. To show this, we give a set-oriented counter example for both S5 and S6. Let $A$, $B$ and $X$ be sets such that $\mathsf{Orthogonal}(X,B)$ and $B \not\subseteq A$. Then obviously $|X \cup A \cup B| > |X \cup A|$, and thus

$$\frac{|X \cap A|}{|X \cup A|} > \frac{|X \cap (A \cup B)|}{|X \cup (A \cup B)|}$$

## 4.3 Cosine coefficient

Next we consider the Cosine coefficient for similarity. The Cosine coefficient stems from the vector model for Information Retrieval. The coefficient is based on the inner vector product, leading to the straightforward style as discussed in the beginning of this section. This coefficient normalizes the intersection $A \cap B$ with the square root of the corresponding product:

$$\mathsf{Cos}(A,B) = \frac{|A \cap B|}{\sqrt{|A| \times |B|}} = \frac{\sum_i a_i b_i}{(\sum_i a_i^2 \ \sum_i b_i^2)^{\frac{1}{2}}} = \frac{A \bullet B}{\|A\|_2 \|B\|_2}$$

in case both $A$ and $B$ are nonempty. If either $A$ or $B$ is empty, we have $\mathsf{Cos}(A,B) = 0$. Finally, $\mathsf{Cos}(\varnothing, \varnothing) = 1$. Note that the number of elements in a set ($|A|$) is related to the euclidian vector length ($\sum_i a_i^2$). This is motivated by $\sum_i a_i^2 = \sum_i a_i$ when $a_i \in \{0,1\}$ for each $i$.

For the weighted case we first notice that $A$ and $B$ are indiscernible iff their enclosed angle equals 0 ($A /\!\!/ B$). Being least similar corresponds with vector orthogonality: $\mathsf{Orthogonal}(A,B) \iff A \perp B$. The base similarity features are easily verified.

A counter example for both S5 and S6 is obtained by taking: $X = (1,0)$, $A = (1,0)$ and $B = (0,1)$, then $\mathsf{Orthogonal}(X,B)$, leading to $\mathsf{Cos}(X,A) = 1$ and $\mathsf{Cos}(X, A \cup B) = 1/\sqrt{2}$ as $A \cup B = (1,1)$.

## 4.4 Dice's coefficient

Next, we consider Dice's similarity coefficient. This coefficient normalizes intersection $A \cap B$ with the average of its constituents:

$$\mathsf{Dice}(A,B) = \frac{2\,|A \cap B|}{|A| + |B|} = \frac{\sum_i \min(a_i, b_i)}{\frac{1}{2}\left(\sum_i a_i + \sum_i b_i\right)} = \frac{\sum_i \min(a_i, b_i)}{\sum_i \mathsf{avg}(a_i, b_i)}$$

in case both $A$ and $B$ are nonempty. If either $A$ or $B$ is empty, then $\mathsf{Dice}(A,B) = 0$. Finally, $\mathsf{Dice}(\varnothing,\varnothing) = 1$. As a consequence:

1. $\mathsf{Identical}(A,B) \iff \sum_i \min(a_i,b_i) = \sum_i \mathsf{avg}(a_i,b_i) \iff \forall_i\,[a_i = b_i] \iff A = B$
2. $\mathsf{Orthogonal}(A,B) \iff A \perp B$

The base similarity features are valid for Dice's coefficient.

A counter example for S5 and S6 is obtained similar as in the case of Jaccard coefficient.

## 4.5 Overlap coefficient

Finally, we consider the Overlap coefficient for similarity. This coefficient normalizes the intersection $A \cap B$ with the minimum cardinality of its arguments:

$$\mathsf{Ovl}(A,B) = \frac{|A \cap B|}{\min(|A|, |B|)} = \frac{\sum_i \min(a_i, b_i)}{\min\left(\sum_i a_i, \sum_i b_i\right)}$$

in case both $A$ and $B$ are nonempty. If either $A$ or $B$ is empty, then $\mathsf{Ovl}(A,B) = 0$. Finally, $\mathsf{Ovl}(\varnothing,\varnothing) = 1$. As a consequence:

1. $\mathsf{Identical}(A,B) \iff \sum_i \min(a_i,b_i) = \min\left(\sum_i a_i, \sum_i b_i\right) \iff \forall_i\,[a_i \le b_i] \vee \forall_i\,[b_i \le a_i] \iff A \le B \vee B \le A$
2. $\mathsf{Orthogonal}(A,B) \iff A \perp B$

The Overlap coefficient does not satisfy the base similarity features. It is easily seen that being identical is not a transitive relation in this case, for example let $A = \{1\}$, $B = \{1,2,3\}$ and $C = \{2\}$. A counter example for S3 is: $A = \{1\}$, $B = \{1,2,3\}$ and $C = \{2\}$.

The monotonicity features are also not satisfied, as a counter example take $X = \{1\}$, $A = \{1\}$ and $B = \{2,3,4,5\}$, then $\mathsf{Ovl}(X,A) = 1$, while $\mathsf{Ovl}(X, A \cup B) = 0.25$. A similar counter example for S6 can be taken.

## 4.6 Weak and strong similarity measures

In [12], a general format for similarity functions is introduced:

$$F(A,B) = \frac{\psi_1(|A \cap B|)}{\psi_2(|A|, |B|, |A \cup B|)}$$

for some functions $\psi_1$ and $\psi_2$. The function $F$ is called a strong similarity function if it satisfies the following rules:

$F_{0a}$**:** $\psi_1$ is a strictly increasing function

$F_{0b}$**:** $\psi_2$ is a strictly increasing function of three variables

$F_1$**:** $0 \leq F(A,B) \leq 1$

$F_2$**:** $F(A,B) = 1 \iff A = B$

$F_3$**:** $F(A,B) = 0 \iff A \cap B = \varnothing$

$F_4$**:** If the denominator of $F$ is constant, then $F$ is strictly increasing with $|A \cap B|$

Note that condition $F_4$ is a consequence of condition $F_{0a}$. For strong similarity functions we have:

1. $\mathsf{Identical}(A,B) \iff A = B$
2. $\mathsf{Orthogonal}(A,B) \iff A \perp B$

From this we conclude that strong similarity functions have the base similarity features. However, strong similarity functions do not satisfy the monotonicity features. For example, Jaccard's coefficient is a strong similarity function without these features.

$F$ is called a weak similarity function if $F_2$ is replaced by:

$F_2'$**:** $F(A,B) = 1 \iff A \subseteq B \vee B \subseteq A$

In this case: $\mathsf{Identical}(A,B) \iff A \leq B \vee B \leq A$. Weak similarity functions do not satisfy the base similarity features, a counter example is the Overlap coefficient.

## 4.7 The information-theoretic approach

In [15] the similarity between objects $A$ and $B$ is studied from the commonalities between these objects ($\mathsf{Common}(A, B)$, and their common description (generality) ($\mathsf{Description}(A, B)$). From a number of assumptions about similarity, the authors derive the following definition for similarity as a fraction of information values of these quantities:

$$\mathsf{IT\text{-}sim}(A, B) = \frac{I(\mathsf{Common}(A, B))}{I(\mathsf{Description}(A, B))}$$

where $I$ is a function to quantify both $\mathsf{Common}(A, B)$ and $\mathsf{Description}(A, B)$. Seeing commonality and generality as events, suggests to use the information value. The information value of an event with probability $p$ is measured as $-\log(p)$. This leads to the Similarity Theorem:

$$\mathsf{IT\text{-}sim}(A, B) = \frac{\log Prob(\mathsf{Common}(A, B))}{\log Prob(\mathsf{Description}(A, B))}$$

In [2], this is further elaborated to obtain an information theoretic similarity measure for documents. In information retrieval, index terms are used to describe the contents of documents. Let $\pi(t)$ be the probability of term $t$ in some document. In terms of tf-idf weighting, the quantity is known as the inverse document frequency, which is weighted by the relevance of the term within the document. This inter-document weight is derived from the term frequency within the document. Let $\mathsf{fr}_A(t)$ be the term weight obtained this way. This leads to the following definition for similarity between documents:

$$\mathsf{IT\text{-}sim}(A, B) = \frac{\sum_t \min(\mathsf{fr}_A(t), \mathsf{fr}_B(t)) \log \pi(t)}{\sum_t \mathsf{avg}(\mathsf{fr}_A(t), \mathsf{fr}_B(t)) \log \pi(t)}$$

Objects may be seen as determined by their frequencies on terms $t$ with $\log \pi(t) > 0$. We use the following notation:

$$\mathsf{IT\text{-}sim}(A, B) = \frac{\sum_i \min(a_i, b_i) p_i}{\sum_i \mathsf{avg}(a_i, b_i) p_i}$$

where $p_i = \log \pi(t)$. This information theoretic motivated similarity measure can be seen as a weighted version of Dice's coefficient. Then

1. $\mathsf{Identical}(A, B) \iff \sum_i \min(a_i, b_i) p_i = \sum_i \mathsf{avg}(a_i, b_i) p_i \iff \sum_i (\min(a_i, b_i) - \mathsf{avg}(a_i, b_i)) p_i = 0 \iff \forall_i [a_i = b_i] \iff A = B$
2. $\mathsf{Orthogonal}(A, B) \iff A \perp B$

The base similarity features are obvious. The monotonicity similarity do not hold. For example, when all weight factors are equal, this similarity measure is equal to Dice's coefficient.

# 5 Conclusions

In this paper the incremental searcher satisfaction model for Information Retrieval has been extended. Starting from a general characterization of cognitive identities, different approaches for the construction of increment functions were studied. During a search process, we need to have a good balance between variety and conciseness. We therefore introduced several primitives in order to switch between exploration and exploitation.

More formally this was established using containment similarity as a basic format for object-set similarity, or, the similarity between a document and a document collection. Two approaches to containment similarity were introduced: reduction and projection. This enabled us to translate the IM-axioms for increment functions into S-axioms for the underlying similarity functions. This is quite an important step, since increment functions are not available everywhere, but (traditional) similarity functions are. Furthermore, using this translation we have confronted our S-axioms with concrete similarity functions, including Inclusion, Jaccard, Cosine, Dice, and Overlap.

We have shown that in the case of incremental IR approaches, the underlying similarity measures should satisfy rather strict requirements. Actually, only the Inclusion function satisfies all requirements.

In future research attention will be focussed on exploring other approaches for the construction of increment functions, and more advanced similarity coefficients will be examined. We are currently working on a context shift, in which our IM-axioms for information searchers, are evaluated in terms of information distributors as well. This results in a dual electronic market, where demand and supply meet.

## References


[1] ACM. Special issue on recommender systems. *Transactions on Information Systems*, 22(1), 2004.

[2] J.A. Aslam and M. Frost. An information-theoretic measure for document similarity. In *SIGIR03, Toronto, Canada*, pages 449–450, July 2003.

[3] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[4] M. Bates. Where should the person stop and the information search start? *Information, Processing and Management*, 26(5):575–591, 1990.

[5] N. J. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, scripts and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3):379–395, 1995.

[6] N.J. Belkin, P.G. Marchetti, and C. Cool. BRAQUE: Design of an interface to support user interaction in information retrieval. *Information Processing & Management*, 29(3):325–344, 1993.

[7] F.C. Berger and P. van Bommel. Personalized Search Support for Networked Document Retrieval Using Link Inference. In R.R. Wagner and H. Thoma, editors, *Proceedings of the 7th International Conference on Database and Expert System Applications (DEXA)*, pages 802–811, Zurich, Switzerland, September 1996. Springer-Verlag.

[8] P.D. Bruza and T.W.C. Huibers. Investigating Aboutness Axioms Using Information Fields. In W.B. Croft and C.J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 112–121, Dublin, Ireland, July 1994. Springer-Verlag.

[9] J.G. Carbonell, Y. Geng, and J. Goldstein. Automated Query-Relevant Summarization and Diversity-Based Reranking. In I. Ferguson, editor, *Proceedings of the IJCAI-97 Workshop on AI and Digital Libraries*, pages 9–14, Nagoya, Japan, August 1997.

[10] Y. Chiarmarella and J.P. Chevallet. About Retrieval Models and Logic. *The Computer Journal*, 35(3):233–242, 1992.

[11] N. Denos, C. Berrut, and M. Mechkour. An image system based on the visualization of system relevance via documents. In *(DEXA 97)*, pages 379–395, 1997.

[12] L. Egghe and C. Michel. Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. *Information Processing and Management*, 39(5):771–807, 2003.

[13] Robert K. France. Weights and Measures: An Axiomatic Model for Similarity Computations.

[14] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artificial Intelligence*, 44:167–207, 1990.

[15] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th International Conference on Machine Learning*, 1998.

[16] R.D. Luce. Semi-order and a theory of utility discrimination. *Econometrica*, 24, 1956.

[17] C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, United Kingdom, 1990.

[18] R. Slowinski and D. Vanderpooten. Similarity relation as a basis for rough approximations, 1995.

[19] I. Tomek and H. Maurer. Helping the user to select a link. *Hypermedia*, 4(2):111–122, June 1992.

[20] A. Tversky. Features of similarity. *Psychological Reviews*, 84(4):327–352, 1977.

[21] H.R. Varian. Economics and search. In *ACM-SIGIR (invited plenary address)*, Berkeley, CA, 1999.

[22] Th. P. van der Weide, T.W.C. Huibers, and P. van Bommel. The Incremental Searcher Satisfaction Model for Information Retrieval. *The Computer Journal*, 41(5):311–318, 1998.

[23] Y.Y. Yao. Measuring Retrieval Effectiveness Based on User Preference of Documents. *Journal of the American Society for Information Science*, 46(2):133–145, 1995.